IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 May 2023, accepted 23 May 2023, date of publication 26 May 2023, date of current version 13 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3280197

RESEARCH ARTICLE

Target Detection Algorithm Incorporating Visual Expansion Mechanism and Path Syndication

QINGYAO LIN[1], RUGANG WANG[1], YUANYUAN WANG[1], FENG ZHOU[1], AND NAIHONG GUO[2]

[1]School of Information Technology, Yancheng Institute of Technology, Yancheng 224051, China
[2]Yancheng Xiongying Precision Machinery Company Ltd., Yancheng 224006, China

Corresponding author: Rugang Wang (wrg3506@ycit.edu.cn)

This work was supported in part by the Jiangsu Graduate Practical Innovation Project under Grant SJCX22_1685, in part by the Major Project of Natural Science Research of Jiangsu Province Colleges and Universities under Grant 19KJA110002, in part by the Natural Science Foundation of China under Grant 61673108, and in part by the Natural Science Research Project of Jiangsu University under Grant 18KJD510010.

ABSTRACT Because the lack of semantic information exchange between characteristic layers, the SSD (Single Shot multibox Detector) algorithm has insufficient detection performance. To address this problem, a detection algorithm called VPE-SSD (Visual Path Enhancement SSD) by incorporating a visual expansion mechanism and path syndication proposed in this paper. Firstly, a visual expansion mechanism is added to the shallow characteristic layer to increase the perceptual field. This enables the semantic information in the shallow layer to be more fully utilized by the network. It can also achieve the purpose of enhancing the expressiveness of the shallow feature layer. Then, the processed deep and shallow characteristic layers are fed into the path syndication module for bi-directional fusion. This improves the global information of the feature layers and generates multi-scale global feature maps. Next, to enhance the detailed information of deep characteristics and improve their expression, the deep characteristic enhancement module is applied to the last three characteristic maps. Finally, using the blended attention module to reduce the negative interference and improve the expression of characteristic maps during target detection. The experimental analysis of the VPE-SSD algorithm is conducted on VOC and COCO, and the mAP is 83.4% and 48.4%. From the result, VPE-SSD algorithm can make better use of the different size characteristic information which helps to improve the performance of the algorithm.

INDEX TERMS Target detection, visual expansion mechanism, path syndication, deep characteristic enhancement, attention mechanism thesaurus.

I. INTRODUCTION

The task of target detection is to use the computer to identify and classify the target object in the input image, and is an integral part of computer vision [1]. Recent years, due to the continuous development of artificial intelligence and deep learning, recognition and detection models developed with Convolutional Neural Networks (CNN) as the cornerstone have achieved more remarkable results in industrial [2], [3], medical [4], [5], transportation [6], [7], [8], image recognition processing [9], [10], [11], [12], [13], and other fields. The main target detection algorithms using deep learning as a framework are Fast R-CNN [14], Faster R-CNN [15], Mask R-CNN [16], YOLO (You look only once) series algorithms [17], [18], [19], [20], [21] and SSD (Single Shot multibox Detection) algorithms [22]. Among the above algorithms, YOLO and SSD algorithms are favored because of their fast detection speed and high detection accuracy. Though the SSD algorithm can extract characteristic maps more suitable for detecting targets of different sizes, it does not fully consider the role of the shallow semantic information and cannot detect effectively in the recognition of the target.

As the SSD algorithm still has some shortcomings, in 2020, Kumar et al. [23] reduced the parameters of

The associate editor coordinating the review of this manuscript and approving it for publication was Yizhang Jiang.

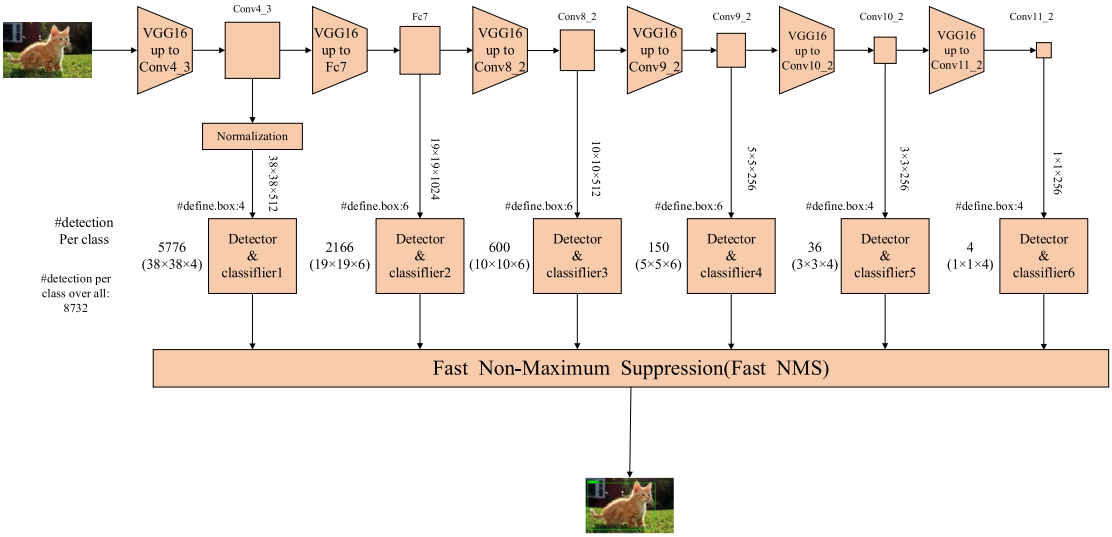


FIGURE 1. Standard SSD algorithm structure diagram.

the model by depth-separable convolution and spatially-separable convolution, allowing the algorithm to achieve desirable results in real-time, but slightly lacking in detection accuracy. Zhaoyuan et al. [24] used Characteristic Pyramid Networks(FPN) to optimize the SSD algorithm, and improved the detection capability of the algorithm by adopting the idea of multi-scale fusion characteristic map. However, the method was limited by the structural design of the characteristic pyramid, and the optimization effect was inadequate. In 2021, Nagrath et al. [25] combined the SSD algorithm and Mobile NetV2 to lighten the algorithm model, which was relatively balanced in terms of detection accuracy and real-time performance, but did not achieve generalization of the scenario. Sun et al. [26] enhanced the target characteristics by capturing the semantic information of deep characteristics, and expanding the scale of the backbone network, thus improving the detection performance of the network. However, this increased the computational power of the algorithm and the hardware burden to a certain extent. Zhang et al. [27] used characteristic mapping to extract global semantic information and proposed an operator called Lightly Expanded Convolution (LMDC), which improved a more accurate input of characteristic information to the detection side. However, the application of shallow characteristics is not sufficient, resulting in poor detection performance in the overlapping part of the target. In 2022, Jiang et al. [28] proposed an improved SSD algorithm that uses deformable convolution to extract target information and improves the detection performance of targets at different scales. However, the use of Res Net50 as the characteristic extraction backbone network, leads to a large computational effort of the algorithm. In 2023, Chintakindi and Hashmi [29] proposed the SSAD algorithm for target detection in autonomous driving using feature fusion and multiscale attention mechanisms, because of the simple and efficient design of SSAD, achieving promising results on different datasets.

To solve the above problems, a target detection algorithm based on SSD fusion visual expansion mechanism and path syndication proposed in this paper. First, the visual expansion mechanism is adopted to expand the perceptual field of shallow characteristics to improve the algorithm's ability to extract target characteristics, and pixel convolution instead of pooling and up-sampling operations is used to reduce the loss of semantic information. Then, the deep characteristics are bi-directionally fused with the shallow characteristics through path syndication to further enhance the semantic information of the shallow characteristics and generate multi-scale global characteristic maps, and the last three characteristic maps are improved by using the deep characteristic enhancement module to strengthen the detail information of the deep characteristics. Finally, the rejection of useless interference information is achieved through a blended attention mechanism to obtain cross-channel and directional location information, which helps the algorithm model to identify and locate detection targets more accurately and improve algorithm performance.

II. RELATED WORK

A. SSD ALGORITHMS

SSD algorithm is a one-stage target detection algorithm based on deep learning proposed by Liu W et al. in 2016, and its structure is shown in Figure.1. VGG16 (Visual Geometry Group network-16) is the backbone network of the SSD algorithm, and the fully connected layer in the YOLO algorithm is replaced by a convolutional layer to obtain different scale characteristic maps. The idea of the anchor frame mechanism is also borrowed in the algorithm to detect targets based on characteristic maps of six sizes. Meanwhile, SSD algorithm combines the advantages of fast detection

speed and accurate candidate localization of the YOLO algorithm, thus achieving higher detection performance.

B. ATTENTION MECHANISM

Inspired by the human visual system, researchers proposed and developed the concept of Attentional Mechanism(AM). In 2014, AM was first proposed as a part of encoder-decoder in the Recurrent Neural Network (RNN) to encode long input utterances [30]. In the last two years, AM has been widely used in deep learning tasks in image recognition, target detection, and natural language processing. The characteristics of the object being recognized differ in importance, which leads to a difference in the importance of each characteristic map in the CNN. Fundamentally, the primary objective of the AM is to identify and extract characteristic information from a large amount of data. Which is more beneficial to task at hand, and it is similar to the human's selective vision mechanism. [31] In 2018, Hu's team [32] proposed the SE-Net (Squeeze-and Excitation Network), which explicitly establishes the interdependencies between characteristic maps on channels and adaptively obtains the importance of different characteristic maps and then updates the weight coefficients. With SE-Net, the team won the ImageNet image classification championship that year. In the same year, Woo's team [33] proposed the CBAM (Convolutional Block Attention Module Network) structure, which extracts the spatial and channel information of characteristic maps using SAM and CAM modules respectively, and fuses them to obtain more stable and reliable characteristics. In 2019, Li's team [34] proposed SK-Net (Selective Kernel Networks), which focuses on the importance among convolutional kernels and considers that the characteristic maps generated from the same image after different convolutional kernels have different importance. In 2020, based on SE-Net, Wang's team [35] proposed ECA-Net (Efficient Channel Attention Network) to achieve cross-channel interaction without dimensionality reduction and a method to adaptively adjust the size of one-dimensional convolutional kernels, which effectively reduces the model complexity while maintaining performance.

The AM is widely used in the field of deep learning mainly relying on its efficient and convenient characteristics, and is also an important means to further improve the effectiveness of deep learning, and is likewise one of the hot spots of current research.

III. ANALYSIS OF THE VPE-SSD ALGORITHM MODEL

The SSD algorithm uses the model to output characteristic maps at different scales to identify the target to be measured, and is able to cover targets of different sizes. However, since the traditional SSD algorithm do not make the most of characteristic information in the shallow layer and do not pay attention to the exchange and fusion of shallow and deep characteristic information, it is deficient in overall performance. In response to this issue, a target detection algorithm called VPE-SSD proposed in this paper that integrates visual expansion mechanism and path syndication, and the overall structure of VPE-SSD is shown in Figure 2. The algorithm consists of four parts: the Visual Expansion Mechanism (VEM), the Path Syndication Module (PSM), the Deep Characteristic Enhancement Module (DFEM) and the Blended Attention Module (BAM). The algorithm is developed based on the SSD algorithm, and after extracting the shallow characteristic layer, the visual expansion mechanism is used to obtain a large perceptual field, so that more effective characteristics can be extracted for algorithm recognition. Meanwhile, the path syndication module is used to bi-directionally fuse the shallow characteristics and deep characteristics to promote the exchange of characteristic information and improve the characteristic extraction capability. When the characteristic information is exchanged, the deep characteristic enhancement module is used to enhance the deep detail characteristic information and improve the expression of deep characteristics. Finally, the blended attention mechanism is used to eliminate the negative effects and background interference in the fusion between different characteristic layers.

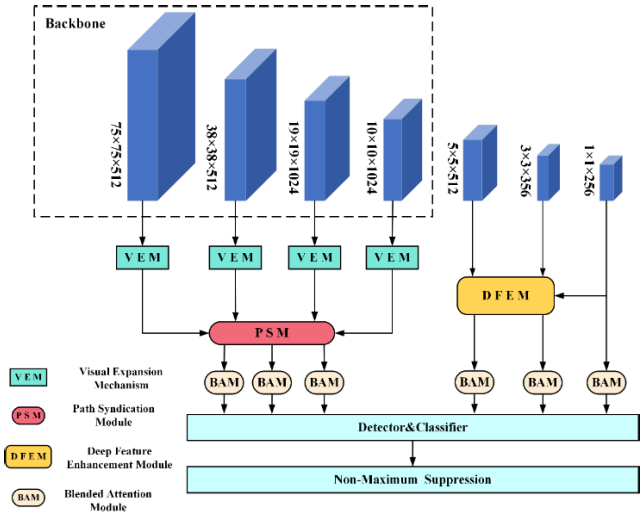


FIGURE 2. VPE-SSD algorithm structure diagram.

A. VISUAL EXPANSION MECHANISM

For the problem of poor detection of shallow targets, a visual expansion mechanism is introduced in the shallow characteristic layer to expand the perceptual field of the model to extract higher-level characteristic information using shallow characteristics. In convolutional neural networks, the size of the input layer corresponding to the output result is determined by the perceptual field; the larger the perceptual field, the more effective information is extracted by the network model, which helps to improve the target detection performance. As shown in Figure3, the visual expansion mechanism adopts a multi-branch structure; each branch uses different convolutional kernels to obtain receptive fields of different sizes while intergrating with the expanded

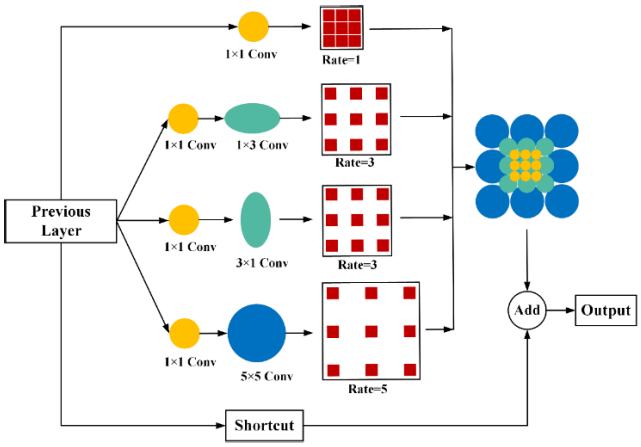


FIGURE 3. Visual expansion mechanism.

convolution to expand the receptive field of the shallow characteristic layer and extract more contextual information. In the visual expansion mechanism, three 1×1 convolutions are used in parallel to adjust the number of channels; then, different convolution operations are used to obtain the characteristic maps; next, expansion convolutions with different expansion rates are used in series in each branch to adjust the perceptual field of the characteristic maps; finally, human visual perception is simulated by concatenate and convolution to obtain the most favorable characteristic maps for target detection. Being close to the human visual perception system, the visual expansion mechanism can obtain discriminative characteristics without increasing the computation amount too much while satisfying the overall real-time requirement of the algorithm.

B. PATH SYNDICATION MODULE

Since the shallow characteristic layer of the SSD algorithm contains only the semantic information of this characteristic layer and lacks global information. For this issue, this paper uses the path syndication module to bi-directionally fuse the deep characteristics with the shallow characteristics to enhance the global information in the shallow characteristic layer and generate a multi-scale global characteristic map, thus improving the detection performance. The overall structure of the path syndication module is presented in Figure 4. Considering that it is not enough to obtain shallow characteristics by Conv4_3 layer alone, which does not express them completely, Conv3_3 layer is added to form a four-way parallel bi-directional path aggregation module to provide more shallow characteristic information for the overall algorithm model. First, Conv3_3, Conv4_3, Conv5_3, and Fc7 in parallel go through the blended attention module and the convolution part with a convolution kernel of 3 to realize interference filtering and characteristic extraction. The convolved Fc7 layer is sub-pixel convolved and processed by the attention module again, then it is stacked with the initial processed Conv5_3 for Concat operation, and the

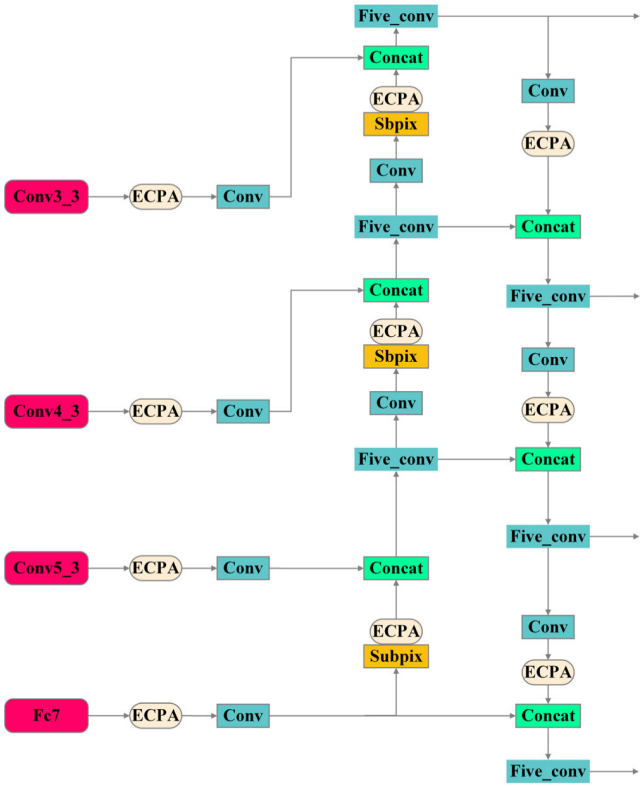


FIGURE 4. Path syndication module.

stacked characteristic layers are further extracted to complete information cross-fertilization of different feature layers. Next, Conv4_3 and Conv3_3 are added to form a complete reverse transmission channel for characteristic exchange from deep layers to shallow layers by the same operation. Subsequently, the fused shallow layer characteristics are convolved and mixed with the attention module to perform the final characteristic extraction and deblending of the characteristic layer, and the characteristics are passed through the shallow to the deep layer to form a characteristic fusion channel for characteristic output. The two-way characteristic exchange channel can effectively facilitate the characteristic information exchange and improve the characteristic extraction capability.

C. CHARACTERISTIC ENHANCEMENT MODULE

Because the low resolution of the deep characteristic layer makes the network model perceive the characteristic details poorly, in order to solve this problem and improve the detection accuracy, this paper also performs characteristic enhancement on the last three characteristic layers of the algorithm. The deep characteristic enhancement module is demonstrated in Figure 5, where the three deep characteristic layers, Conv9_2, Conv10_2 and Conv11_2, are first convolved with a kernel of 3×3 in parallel, then Conv9_2 is sub-pixel convolved to compensate for the information lost in up-sampling, the number of channels is

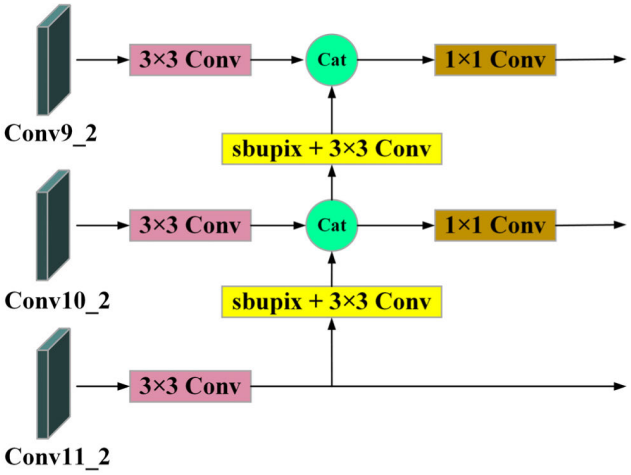


FIGURE 5. Deep characteristic enhancement module.

adjusted to be consistent with Conv10_2 by convolution for stacking operation, and then the result in output by 1×1 convolution. Similarly, the stacked characteristic map is sub-pixel convolved, the number of channels is adjusted to be stacked with Conv9_2, and then the result is output by 1×1 convolution. The purpose of using the deep characteristic enhancement module is to improve the correlation between different characteristic layers and extract richer characteristic information. This structure is suitable for solving the problem of local ambiguity between characteristics and helps to better classify the targets to be detected.

D. BLENDED ATTENTION MODULE

The algorithm will be affected by background interference and useless information in the detection process, also, the channel and location information between each characteristic map differ, and cannot reflect the correlation and importance of each other. This problem will bring a negative impact on the algorithm model and result in poor detection results. To solve this problem, this paper designs a blended attention mechanism that combines channel attention and location attention, as shown in Figure 6. The rich location information in shallow characteristics can be used to extract the dependencies of target location characteristics, while the rich semantic information in deep characteristics can better reflect the importance of target characteristics. In, the input characteristic map is sent to the channel attention module for processing by global average pooling, then 1D convolution instead of the fully connected layer is directly applied to obtain better cross-channel information acquisition with less overhead, and, multiply and fuse the processed characteristic maps with the input characteristic maps. Next, the processed characteristic maps are input to the position attention module in series to process the position information, and three characteristic maps P_{i1}, P_{i2}, and P_{i3} are generated after three 1×1 convolutions in parallel; then, P_{i1} is reshaped and transposed to obtain the matrix P_{i1}^{T}, and P_{i2} and P_{i3} are reshaped to obtain P'_{i2} and P'_{i3}. Subsequently, the correlation matrix D is obtained by multiplying P_{i1}^{T} and P_{i2}, and it is reshaped to obtain the characteristic map, then, average pooling with Sigmoid activation is added to achieve the attention matrix A. Finally, A and P'_{i3} is multiplied element by element, and the result is added it to the characteristic map P_i to obtain the final characteristic map P_i^{out} P_i^out with location characteristic information. The process of whole attention module is expressed as

$$\begin{cases} P_i = Conv1D(Gavgpool(F_i)) \otimes F_i \\ P_{i1} = Conv(P_i) \\ P_{i2} = Conv(P_i) \\ P_{i3} = Conv(P_i) \\ P_{i1}^{T} = Tran(\mathrm{Re}(P_{i1})) \\ P'_{i2} = \mathrm{Re}(P_{i2}) \\ P'_{i3} = \mathrm{Re}(P_{i3}) \\ D = P_{i1}^{T} \otimes P'_{i2} \\ A = Sig(Avg(\mathrm{Re}(D))) \\ P_i^{out} = (A \oplus P'_{i3}) \oplus P_i \end{cases} \tag{1}$$

In Equation 1, $Conv1D()$ denotes 1D convolution, $Gavgpool()$ denotes the global average pooling function, $Conv()$ denotes 1×1 convolution with the relu activation layer, $Re()$ denotes the reshape operation, $Tran()$ denotes the transpose operation, $\otimes$ denotes the element-by-element multiplication operation, $Avg()$ denotes the average pooling function, $Sig()$ denotes the Sigmoid activation function, and $\oplus$ denotes the element-by-element summation operation.

IV. EXPERIMENTS AND ANALYSIS OF RESULTS

A. EXPERIMENTAL SETUP

To test whether the effectiveness of the algorithm proposed in this paper, two general datasets, PASCAL VOC2007+2012 and MS COCO, are selected for validation analysis in experiments. The targets in the VOC are divided into 20 classes, including 16551 training images with 40058 targets, 8333 validation images with 20148 targets, and 4952 test images. The targets are subdivided into 80 categories in COCO2017, including 118287 training images, 5000 validation images, and 40,670 test images. Both of the above datasets are suitable for application in performance testing of algorithms

The experimental platform is NVIDIA GeForce RTX3090 GPU with 24GB memory. The algorithm is implemented in the TensorFlow2.4 framework using Python3.7.9 as the compiler. In the model training process, SGD (Stochastic gradient descent) is used as the optimizer, and the specific parameters are listed in Table 1.

When evaluating the algorithm's performance, detection accuracy and detection speed,i.e., mAP (mean Average Precision) and FPS (Frames Per Second), are taken as evaluation metrics. Specifically, mAP is the average of all categories of AP, and AP is defined in Equation 2, where p(r) denotes the recall and accuracy curves, and FPS

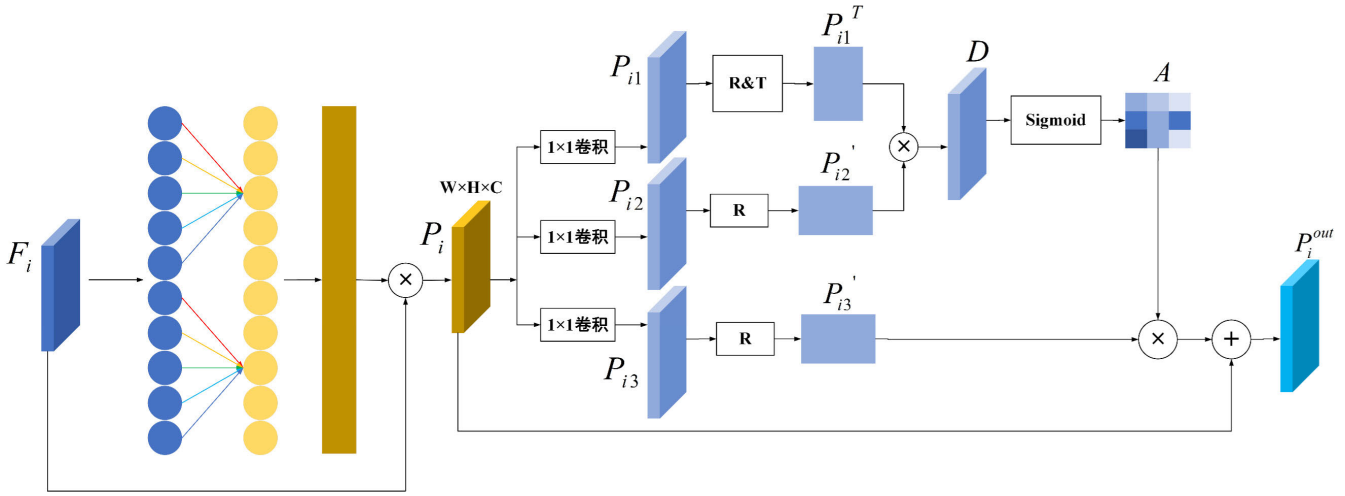


FIGURE 6. Blended attention module.

TABLE 1. Experimental parameter setting.

Parameter	Value
Momentum	0.937
Learning_rate	0.002
Weight_decay	0.0005
Epoch	300
Batch_size	32

indicates the number of images the algorithm can process per second.

$$AP = \int_0^1 p(r)dr \tag{2}$$

The loss curve of the VPE-SSD algorithm is shown in Figure 7, when the algorithm is applied on the VOC dataset, the loss value decreases from 18 and it converges to around 3.5 finally, as shown in Figure 7(a). When the algorithm is trained on COCO, the loss value decreases from 19 and it converges around 3.3 finally, as shown in Figure 7(a).

B. ANALYSIS OF RESULTS

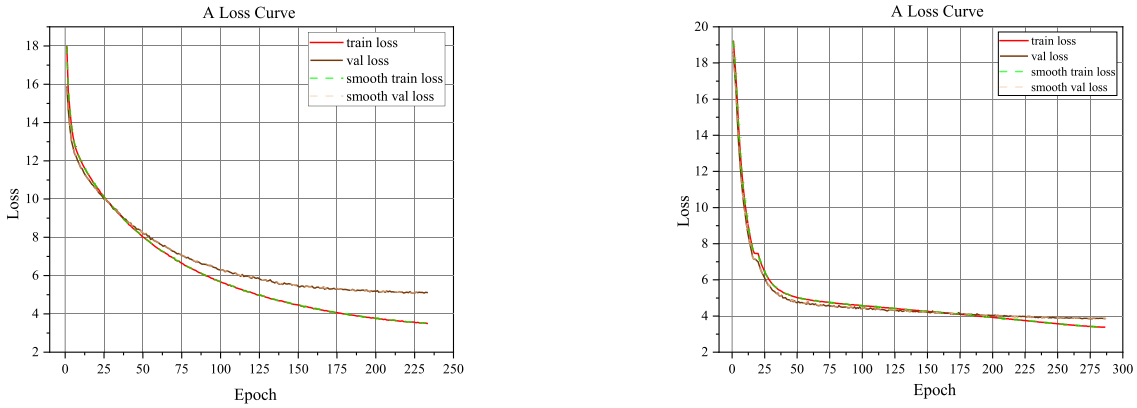
To verify the effectiveness of the improvements of the algorithm in this paper, this section mainly compares the performance of the proposed VPE-SSD algorithm with the target detection algorithms based on CNNs in recent years.

1) EXPERIMENTAL RESULTS OF VOC

Table 2 lists the performance comparison between VPE-SSD and the target detection algorithms of recent years on VOC. The results are all derived on the VOC07+12 dataset. The mAP results are the mean values of the various types of accuracy detected with the intersection ratio of positive and negative sample areas of 0.5. The results in Table 2 indicate that the VPE-SSD algorithm can achieve a detection accuracy of 81.2% and detection speed of 23.6 FPS for an input size of 300 × 300. Compared to the RFB and SSD algorithms, the detection accuracy is improved by 0.7% and 4% respectively; compared to SSD-based optimization algorithms such as DSSD, MDSSD, DF-SSD, RSSD, FSSD, and SEFN, the detection accuracy is improved by 2.6%, 2.6%, 2.3%, 2.7%, 2.4%, and 1.6%, respectively. At an input size of 512 × 512, the VPE-SSD algorithm can achieve a detection accuracy of 83.4% and detection speed of 16.1 FPS. Compared to the RFB and SSD algorithms, the detection accuracy is improved by1.2% and 4.9% respectively; compared to SSD-based optimization algorithms such as DSSD, MDSSD, RSSD, FSSD, ESSD, and SEFN, the detection accuracy is improved by 1.9%, 2.4%, 2.6%, 2.5%, 1.3%, and 2.2%, respectively. Since images inevitably have smaller targets to be measured, different sizes are used in Table 2 to help improve the detection of small objects to improve the comprehensiveness of the algorithm. So we choose two different sizes as input, and from the results in Table 2 we can see that the larger size has a more significant improvement than the smaller size. Although VPE-SSD has improved the detection accuracy compared with the above algorithms and can meet the requirements of real-time detection, there are still some shortcomings.

To verify the effectiveness of each module of the VPE-SSD algorithm, ablation experiments are conducted on the VOC07+12 dataset, and the models, i.e., VEM, PSM, DFEM and BAM are added to the basic model of the SSD algorithm one by one. The effectiveness of these modules is analyzed by comparing the experimental results listed in Table 3.

In verifying that the VEM is valid, the SSD is used as the basis, and VEM was added to four characteristic layers, i.e., Conv3_3, Conv4_3, Conv5_3, and Fc7, to enhance



(a) Training loss curve of the VOC dataset.

(b) Training loss curve of the COCO dataset.

FIGURE 7. VPE-SSD model training loss curve.

TABLE 2. mAP of different algorithms on VOC dataset.

Methods	Backbone	Size	mAP(%)	FPS
Faster R-CNN[15]	VGG16	1000×600	73.2	7.0
Faster R-CNN[15]	ResNet101	1000×600	76.4	2.4
SSD300[22]	VGG16	300×300	77.2	46.0
SSD512[22]	VGG16	512×512	78.5	19.0
DSSD321[36]	ResNet-101	321×321	78.6	9.5
DSSD513[36]	ResNet-101	513×513	81.5	5.5
MDSSD300[37]	VGG16	300×300	78.6	32.2
MDSSD512[37]	VGG16	512×512	81.0	14.5
DF-SSD[38]	DenseNet-S-32-1	300×300	78.9	11.6
SEFN300[39]	VGG16	300×300	79.6	55.0
SEFN512[39]	VGG16	512×512	81.2	30.0
RSSD300[40]	VGG16	300×300	78.5	35.0
RSSD512[40]	VGG16	512×512	80.8	16.6
FSSD300[41]	VGG16	300×300	78.8	**65.8**
FSSD512[41]	VGG16	512×512	80.9	35.7
RFB300[42]	VGG16	300×300	80.5	83.0
RFB512[42]	VGG16	512×512	82.2	38.0
ESSD[43]	VGG16	512×512	82.1	15.7
VPE-SSD300	**VGG16**	**300×300**	**81.2**	23.6
VPE-SSD512	**VGG16**	**512×512**	**83.4**	16.1

TABLE 3. Ablation experiments for each module.

VEM	PSM	DPEM	BAM	VOC 07+12 mAP/%
				72.4
√				75.3
√	√			78.6
√	√	√		79.2
√	√	√	√	81.2

the perception of target characteristics by increasing the perceptual field of the algorithm. The results in Table 3 indicate that the mAP is increased by 2.9% after adding the VEM module, indicating that VEM can provide the algorithm with useful characteristic information for detecting targets.

In verifying that the PSM is valid, the module is added to the four characteristic layers of Conv3_3, Conv4_3, Conv5_3, and Fc7 after the VEM module for path syndication. The results show that the mAP is improved by 3.3%, indicating that the PSM module can better fuse the semantic information of deep and shallow layers, thus improving the detection performance of the algorithm.

In verifying that the DPEM is valid, characteristic enhancement is performed on three deep characteristic layers, i.e., Conv9_2, Conv10_2 and Conv11_2, based on the addition of the first two modules. The mAP is improved by 0.6%, indicating the usefulness of DPEM in improving the correlation between different characteristic layers.

In verifying that the BAM is valid, BAM is added after the three modules mentioned above, and the mAP is improved by 2%. The result indicates that BAM can eliminate the undesirable effects of background and characteristic fusion,

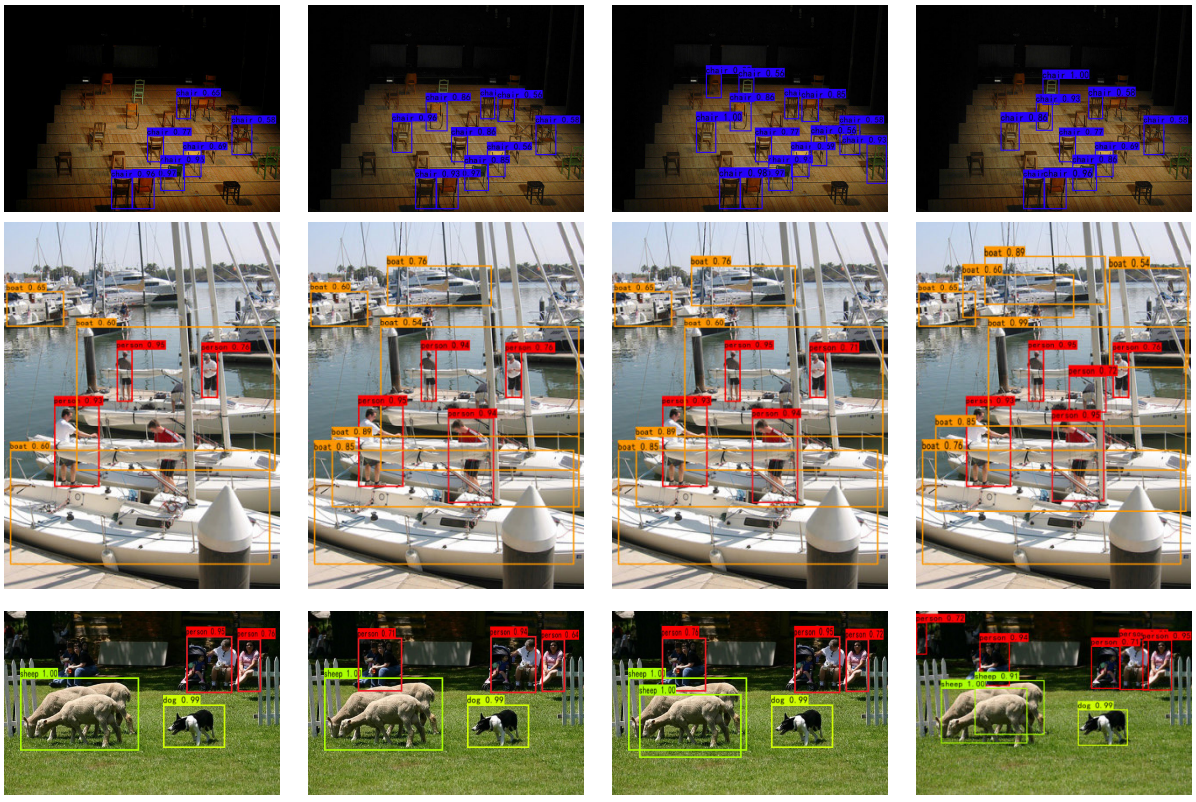


FIGURE 8. Graph of qualitative experimental results of different algorithms.

TABLE 4. Performance comparison of the VPE-SSD algorithm and other algorithms on the MS COCO dataset.

Methods	Backbone	Avg. precision, IoU			Avg. precision, area			Avg. recall, area		
		IOU=0.5:0.95	IOU=0.5	IOU=0.75	Area: S	Area: M	Area: L	Area: S	Area: M	Area: L
SEFN512[39]	VGG16	33.7	54.7	35.6	**19.2**	38.0	47.3	**29.1**	**52.5**	63.2
SSD512[22]	VGG16	27.7	46.4	26.7	10.9	31.8	43.5	16.5	46.6	60.8
FSSD512[41]	VGG16	31.8	52.8	33.5	14.2	35.1	45.0	22.3	49.9	62.0
DF-SSD[40]	DenseNet-S-32-1	29.5	50.7	31.3	9.8	31.1	46.5	17.3	46.8	64.4
DSOD300[44]	DS/64-192-48-1	29.3	47.3	30.6	9.4	31.5	47.0	16.7	47.1	65.0
DSSD513[36]	ResNet-101	33.2	53.3	35.2	13.0	35.4	51.1	28.9	43.5	46.2
RFB512[42]	VGG16	34.4	55.7	36.4	17.6	37.0	47.6	27.3	52.3	65.4
VPE-SSD	**VGG16**	**48.4**	**76.6**	**52.6**	17.4	36.7	**56.2**	26.1	48.4	**66.2**

which helps to balance the information between characteristic maps.

2) QUALITATIVE EXPERIMENTS

Figure 8 compares the detection effect of VPE-SSD, SSD, SEFN, and ESSD algorithms on VOC. From the results displayed, the performance of the SSD is insufficient, and there is an obvious phenomenon of missing detection. The SEFN algorithm and ESSD algorithm are adjusted and optimized based on SSD algorithm, and although they alleviate the missing detection phenomenon to a certain extent, they are not effective in detecting overlapping targets and incomplete targets. By incorporating the visual expansion mechanism and path syndication, our proposed target detection algorithm can deal with these situations more effectively and shows a more desirable detection performance.

3) EXPERIMENTAL RESULTS OF COCO

To further demonstrate the performance advantages of the proposed VPE-SSD algorithm for target detection, this paper further compares it with other algorithms on the MS COCO dataset, and the results are shown in Table 4. It can be seen that the VPE-SSD algorithm improves the detection accuracy and recall to varying degrees, compared to algorithms

such as SEFN512, SSD512, FSSD512, DF-SSD, DSOD300, DSSD513, and RFB512. In Table 4, the results indicate that the proposed VPE-SSD algorithm has better performance in target detection.

V. CONCLUSION

To solve the problem of poor target detection performance due to inadequate semantic information and lack of information exchange between characteristic layers, a algorithm called VPE-SSD proposed in this paper that incorporates the visual expansion mechanism and path syndication. Four components are designed to optimize performance of the SSD algorithm: the visual expansion mechanism, the path syndication module, the deep characteristic enhancement module, and the blended attention module. Firstly, a visual expansion mechanism is added to the shallow characteristic layer to increase the perceptual field, so that the semantic information in the shallow layer can be more fully utilized by the network for the purpose of enhancing the expression of the shallow characteristics; then, the processed deep and shallow characteristic layers are fed into the path aggregation module for bi-directional fusion to improve the global information of the characteristic layers and generate multi-scale global characteristic maps. Next, to improve the expression of deep characteristics, the last three characteristic maps are enhanced using the deep characteristic enhancement module to enhance the detailed information of deep characteristics. Finally, using the blended attention module to reduce the negative interference and improve the expression of characteristic maps during target detection.

The result of the VPE-SSD algorithm was conducted on the VOC and COCO datasets. The mAP of the algorithm in this paper was 83.4% and 48.4%, respectively, and the results indicate that the VPE-SSD algorithm achieves better detection performance. The VPE-SSD algorithm proposed in this paper can effectively improve the utilization of semantic information and enhance the information exchange between feature layers to improve the detection accuracy of the algorithm. However, the size of the VPE-SSD algorithm is not fully considered, which leads to a slightly inferior processing speed. Our future work will study how to improve the speed of the algorithm without reducing the accuracy.

DATA AVAILABILITY

The data used to support the findings of this study are available from the corresponding author upon request.

CONFLICTS OF INTEREST

There is no conflict of interest regarding the publication of this articles.

REFERENCES

[1] X. Xiao, B. Wang, L. Miao, L. Li, Z. Zhou, J. Ma, and D. Dong, "Infrared and visible image object detection via focused feature enhancement and cascaded semantic extension," *Remote Sens.*, vol. 13, no. 13, p. 2538, Jun. 2021.

[2] T. Gao, X. Zhang, and B. Li, "Review on the application of deep learning in helmet wearing detection," *Comput. Eng. Appl.*, to be published.

[3] C. J. L. Bin, "Traffic intersection target detection based on improved SSD lightweight," *Transducer Microsyst. Technol.*, vol. 41, no. 10, pp. 117–121, 2022, doi: 10.13873/J.1000-9787(2022)10-0117-05.

[4] H. Di, L. Lixin, L. Yujie, and X. Feng, "Object detection of pneumonia images based on deep learning," *Chin. J. Biomed. Eng.*, vol. 41, no. 4, pp. 443–451,2022.

[5] Y. Ming, T. Dapeng, and P. Yuanyuan, "Texture-less object detection method for industrial components picking system," *J. Image Graph.*, vol. 27, no. 8, pp. 2418–2429,2022.

[6] C. Liang, B. Jie, and H. Libo, "Multi-target detection based on camera and radar characteristic fusion networks," *Trans. Beijing Inst. Technol.*, vol. 42, no. 3, pp. 318–323,2022, doi: 10.15918/j.tbit1001-0645.2021.164.

[7] W. Liuyi, S. Wen'fai, and L. Xinshan, "Review of computer-aided diagnosis of bronchiectasis with CT images," *Comput. Eng. Appl.*, vol. 57, no. 11, pp. 11–20, 2021.

[8] Z. Sun, X. Leng, Y. Lei, B. Xiong, K. Ji, and G. Kuang, "BiFA-YOLO: A novel YOLO-based method for arbitrary-oriented ship detection in high-resolution SAR images," *Remote Sens.*, vol. 13, no. 21, p. 4209, Oct. 2021.

[9] H. Guo, Q. Shi, A. Marinoni, B. Du, and L. Zhang, "Deep building footprint update network: A semi-supervised method for updating existing building footprint from bi-temporal remote sensing images," *Remote Sens. Environ.*, vol. 264, Oct. 2021, Art. no. 112589.

[10] M. Liu, Q. Shi, J. Li, and Z. Chai, "Learning token-aligned representations with multimodel transformers for different-resolution change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4413013.

[11] W. Debin and L. Xiaonan, "Small object detection algorithm fusing vision mechanism and multi-scale characteristics," *Telecommun. Eng.*, to be published.

[12] Q. Lin, S. Li, R. Wang, Y. Wang, F. Zhou, Z. Chen, and N. Guo, "Research on small target detection technology based on the MPH-SSD algorithm," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–13, Nov. 2022, doi: 10.1155/2022/9654930.

[13] X. Liu, R. Wu, and R. Wang, "Bearing fault diagnosis based on particle swarm optimization fusion convolutional neural network," *Frontiers Neurorobot.*, to be published, doi: 10.3389/fnbot.2022.104496.

[14] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[15] S. Ren, K. He, and R. Girshick, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015.

[16] K. He, G. Gkioxari, and P. Dollar, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.

[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[18] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[19] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[20] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, Y. Li, B. Zhang, Y. Liang, L. Zhou, X. Xu, X. Chu, X. Wei, and X. Wei, "YOLOv6: A single-stage object detection framework for industrial applications," 2022, *arXiv:2209.02976*.

[21] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, *arXiv:2207.02696*.

[22] W. Liu, D. Anguelov, and D. Erhan, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.

[23] A. Kumar, Z. J. Zhang, and H. Lyu, "Object detection in real time based on improved single shot multi-box detector algorithm," *EURASIP J. Wireless Commun. Netw.*, vol. 2020, no. 1, pp. 1–18, Dec. 2020.

[24] Y. Zhaoyuan and Z. Jianli, "Object detection algorithm DFSSD based on automatic driving scene," *Comput. Eng. Appl.*, vol. 56, no. 16, p. 9, 2020.

[25] P. Nagrath, R. Jain, A. Madan, R. Arora, P. Kataria, and J. Hemanth, "SSDMNV2: A real time DNN-based face mask detection system using single shot multibox detector and MobileNetV2," *Sustain. Cities Soc.*, vol. 66, Mar. 2021, Art. no. 102692.

[26] P. Sun, Y. Zhao, and S. Zhu, "An approach to improve SSD through mask prediction of multi-scale feature maps," *Pattern Anal. Appl.*, vol. 24, no. 3, pp. 1357–1366, Aug. 2021.

[27] X. Zhang, H. Xie, Y. Zhao, W. Qian, and X. Xu, "A fast SSD model based on parameter reduction and dilated convolution," *J. Real-Time Image Process.*, vol. 18, no. 6, pp. 2211–2224, Dec. 2021.
[28] J. Chen, Q. Yongming, and Y. Xingtian, "Target detection method based on deformable convolution improved SSD algorithm," *Electron. Meas. Technol.*, vol. 45, no. 16, pp. 116–122, 2022, doi: 10.19651/j.cnki.emt.2209283.
[29] B. M. Chintakindi and M. F. Hashmi, "SSAD: Single-shot multi-scale attentive detector for autonomous driving," *IETE Tech. Rev.*, pp. 1–13, Feb. 2023.
[30] L. Bin, L. Quan, and X. Jin, "Target specific emotion analysis based on multi attention convolutional neural network," *Comput. Res. Develop.*, vol. 54, no. 18, pp. 1724–1735, 2017.
[31] Z. Feng, H. Xiaofeng, and W. Lin, "From situation cognition to situation intelligent cognition," *J. Syst. Simul.*, vol. 30, no. 13, pp. 761–771, 2018.
[32] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[33] S. Woo, J. Park, and J. Y. Lee, "CBAM: Convolutional block attention module," 2018, *arXiv:1807.06521*.
[34] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 510–519.
[35] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11534–11542.
[36] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: Deconvolutional single shot detector," 2017, *arXiv:1701.06659*.
[37] L. Cui, R. Ma, P. Lv, X. Jiang, Z. Gao, B. Zhou, and M. Xu, "MDSSD: Multi-scale deconvolutional single shot detector for small objects," 2018, *arXiv:1805.07009*.
[38] S. Zhai, D. Shang, S. Wang, and S. Dong, "DF-SSD: An improved SSD object detection algorithm based on DenseNet and feature fusion," *IEEE Access*, vol. 8, pp. 24344–24357, 2020.
[39] H. Chen and H. Luo, "Multi-scale semantic information fusion for object detection," *J. Electron. Inf. Technol.*, vol. 43, no. 7, pp. 2087–2095, 2021.
[40] J. Jeong, H. Park, and N. Kwak, "Enhancement of SSD by concatenating feature maps for object detection," 2017, *arXiv:1705.09587*.
[41] Z. Li and F. Zhou, "FSSD: Feature fusion single shot multibox detector," 2017, *arXiv:1712.00960*.
[42] S. Liu and D. Huang, "Receptive FELD block net for accurate and fast object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 385–400.
[43] Z. Weiliang and C. Xiuhong, "SSD object detection algorithm with cross-layer fusion and receptive field amplification," *Comput. Sci.*, to be published. [Online]. Available: http://kns.cnki.net/kcms/detail/50.1075.TP.20221109.1716.026.html
[44] Z. Shen, Z. Liu, J. Li, Y. Jiang, Y. Chen, and X. Xue, "DSOD: Learning deeply supervised object detectors from scratch," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1937–1945.

QINGYAO LIN received the B.S. degree from the Yancheng Institute of Technology, Yancheng, China, in 2021, where he is currently pursuing the M.Eng. degree. His current research interests include computer vision technology and intelligent control systems and signal detection.

RUGANG WANG received the B.S. degree from the Wuhan University of Technology, Wuhan, China, in 1999, the M.S. degree from Jinan University, Guangzhou, China, in 2007, and the Ph.D. degree from Nanjing University, Nanjing, China, in 2012. He is currently a Professor with the College of Information Engineering, Yancheng Institute of Technology, Yancheng, China. His research interests include optical communication networks, novel and key devices for optical communication systems, and image processing technology.

YUANYUAN WANG, photograph and biography not available at the time of publication.

FENG ZHOU received the B.S. and M.S. degrees from Southeast University, Nanjing, China, in 2004 and 2012, respectively. He is currently pursuing the Ph.D. degree with the Army Engineering University of PLA. Since 2017, he has been an Associate Professor with the College of Information Engineering, Yancheng Institute of Technology, Yancheng, China. His research interests include cooperative communication, computer vision technology, and image processing technology.

NAIHONG GUO, photograph and biography not available at the time of publication.